Essential Feedback Loops

systems and apps, both man and machine

Blue Water / TeraGrid Fault-Tolerance Workshop

Albuquerque, New Mexico

March 19, 2009

Jon Stearley <jrstear@sandia.gov>
(With input from Bob Balance and Sue Kelly)

Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company,
for the United States Department of Energy's National Nuclear Security Administration
under contract DE-AC04-94AL85000.

Questions for speaker:

- What can you tell us about your fault/error situation currently?
  - What rate of errors?
  - What types of errors do you see?
  - How does one know if there is a fault?

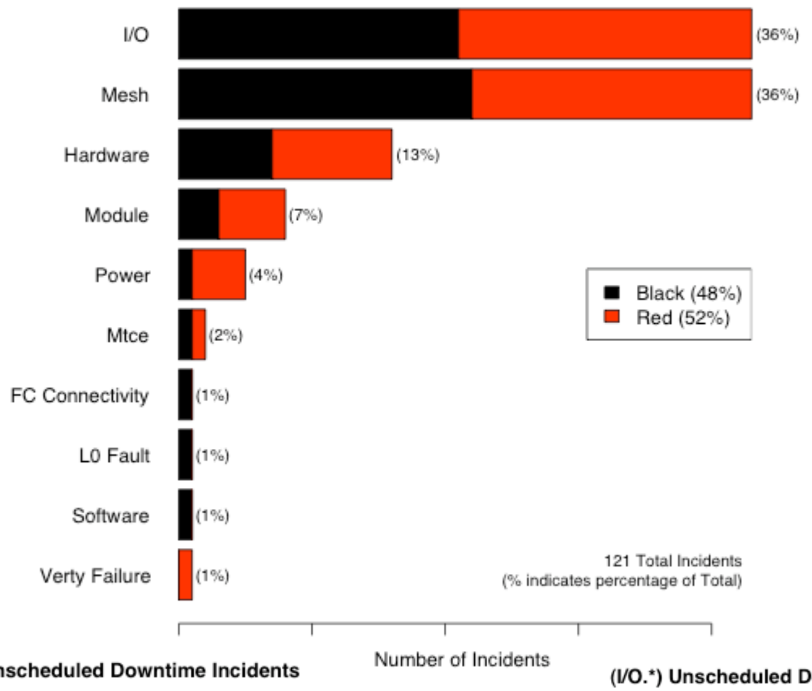
- What are you worried about in the future?
  - What do you think is needed to help?
  - What do you think it is reasonable for apps people to do?

Redstorm (both sides) causes of Unscheduled Downtime (lifetime)

CAUSE	CATEGORY						
	HW	SW	Systemic	RAS	ENV	PROC	UNK
Cache Parity Error	4	0	0	0	0	0	0
Color Change	0	0	0	0	0	0	0
FC Connectivity	1	0	0	0	0	0	0
Hardware	49	0	0	0	0	0	0
I/O: Hang	0	4	0	0	0	0	0
I/O: Indeterminate Error	0	0	9	0	0	0	0
I/O:Channel error	0	0	14	0	0	0	0
I/O:DDN:	83	0	0	0	0	0	0
I/O:DDN:Offline	7	0	0	0	0	0	0
I/O:DDN:Reboot	14	0	0	0	0	0	0
I/O:Hang	0	9	0	0	0	0	0
I/O:LSI:	4	0	0	0	0	0	0
I/O:Lustre	0	31	0	0	0	0	0
I/O:Qlogic HBA	13	0	0	0	0	0	0
L0 Fault	0	0	0	14	0	0	0
Mesh	66	0	0	0	0	0	0
Mesh:Cable	5	0	0	0	0	0	0
Mesh:Deadlock	39	0	0	0	0	0	0
Mesh:FIFO Overrrun	0	0	3	0	0	0	0
Mesh:LUT error	0	0	8	0	0	0	0
Mesh:Link Failed	26	0	0	0	0	0	0
Mesh:Link Inactive	104	0	0	0	0	0	0
Module:Powerdown	13	0	0	0	0	0	0
Module:Powerdown:VRM	3	0	0	0	0	0	0
Module:Powerdown:Verty	33	0	0	0	0	0	0
Mtce	0	0	0	0	0	2	0
Portals	0	23	0	0	0	0	0
Power:Cabinet:EPO	11	0	0	0	0	0	0
Power:Cable	1	0	0	0	0	0	0
Power:Outage	0	0	0	0	13	0	0
Procedural	0	0	2	0	0	7	0
Procedural:Operator Error	0	0	0	0	0	3	0
SDB Crashed	0	0	4	0	0	0	0
Software	0	23	0	0	0	0	0
TCP/IP:CRC	0	4	0	0	0	0	0
Test:System	0	0	0	0	0	0	0
Unknown	0	0	0	0	0	0	11
Upgrade:HW	0	0	0	0	0	0	0
Upgrade:SW	0	4	0	0	0	0	0
Verty Failure	4	0	0	0	0	0	0
TOTAL	480	98	40	14	13	12	11
%	72%	15%	6%	2%	2%	2%	2%

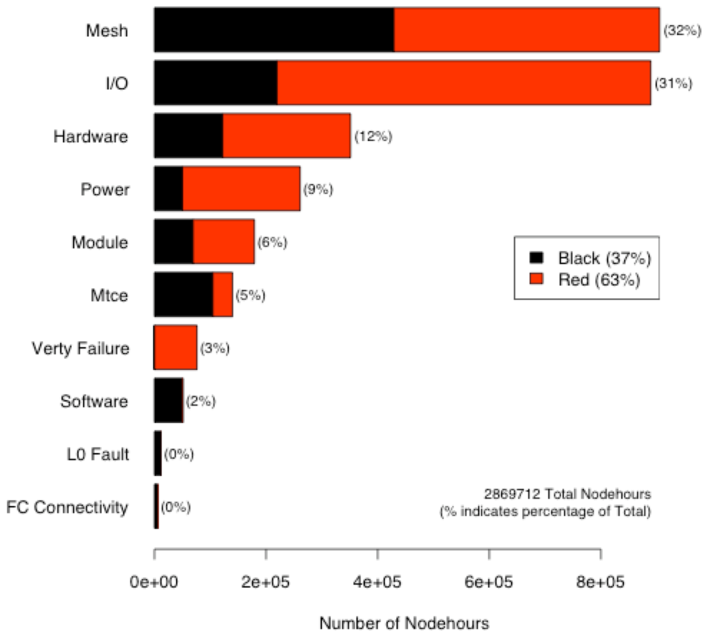
Primary Causes of Unscheduled Downtime Incidents

Red Storm: 07/01/08 through 12/31/08



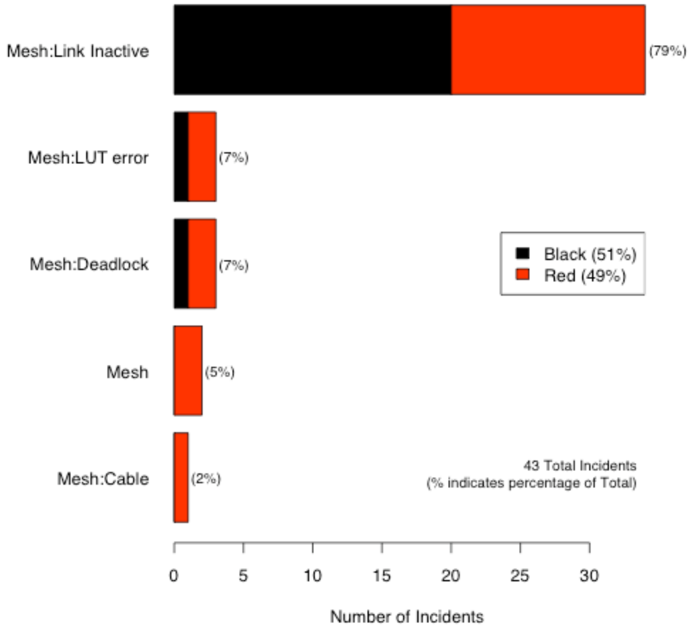
Primary Causes of Unscheduled Downtime Nodehours

Red Storm: 07/01/08 through 12/31/08



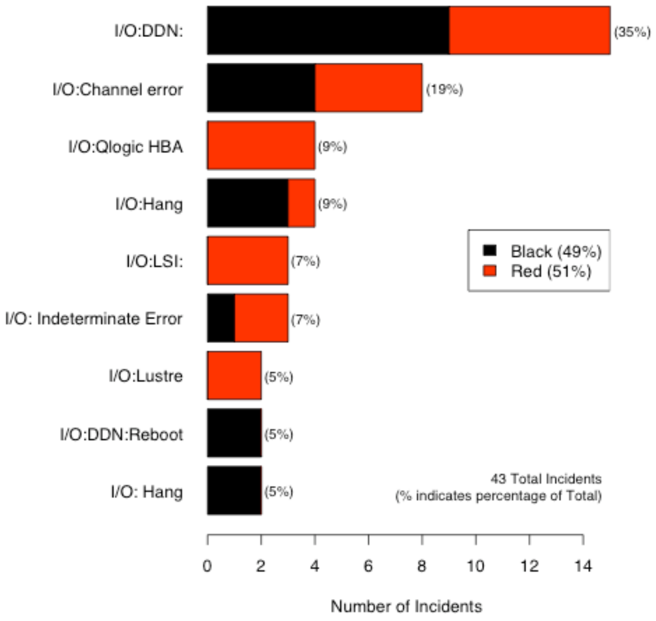
(Mesh.*) Unscheduled Downtime Incidents

Red Storm: 07/01/08 through 12/31/08



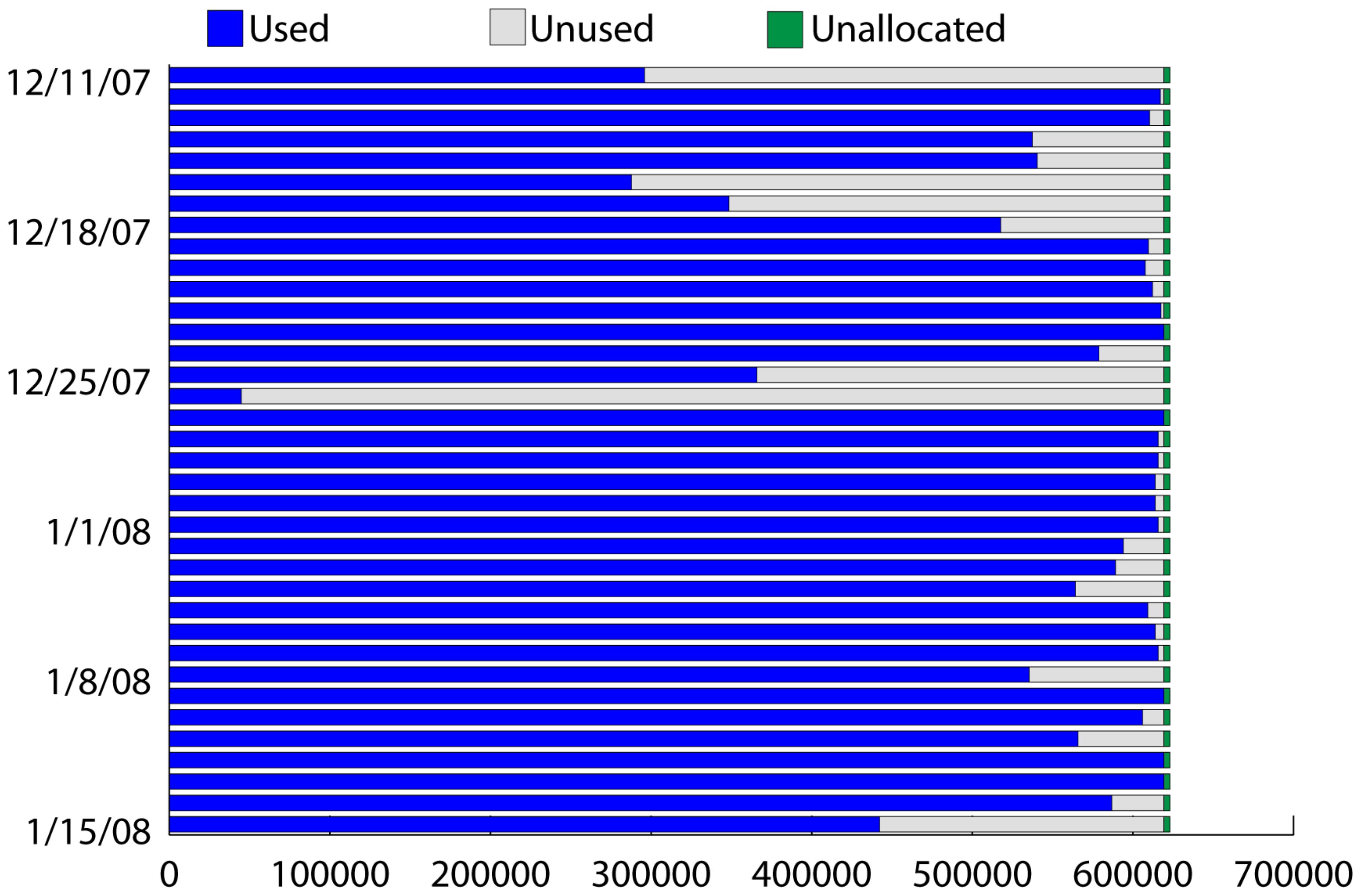
(I/O.*) Unscheduled Downtime Incidents

Red Storm: 07/01/08 through 12/31/08

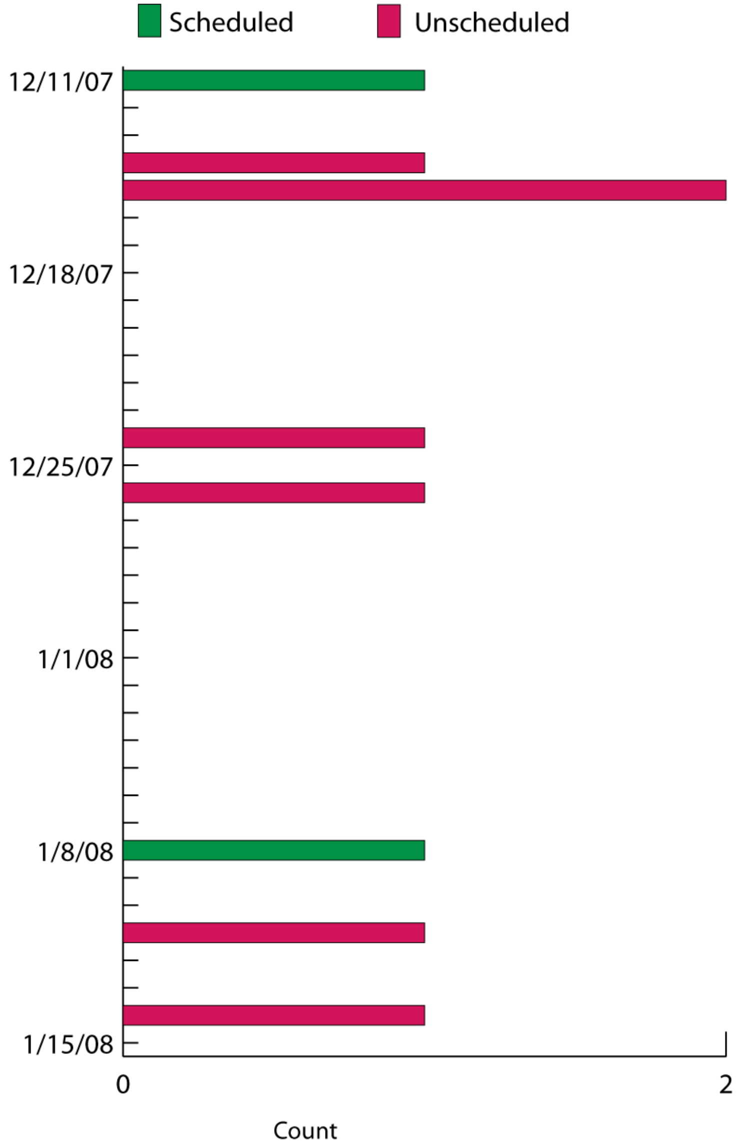


Sandia National Laboratories

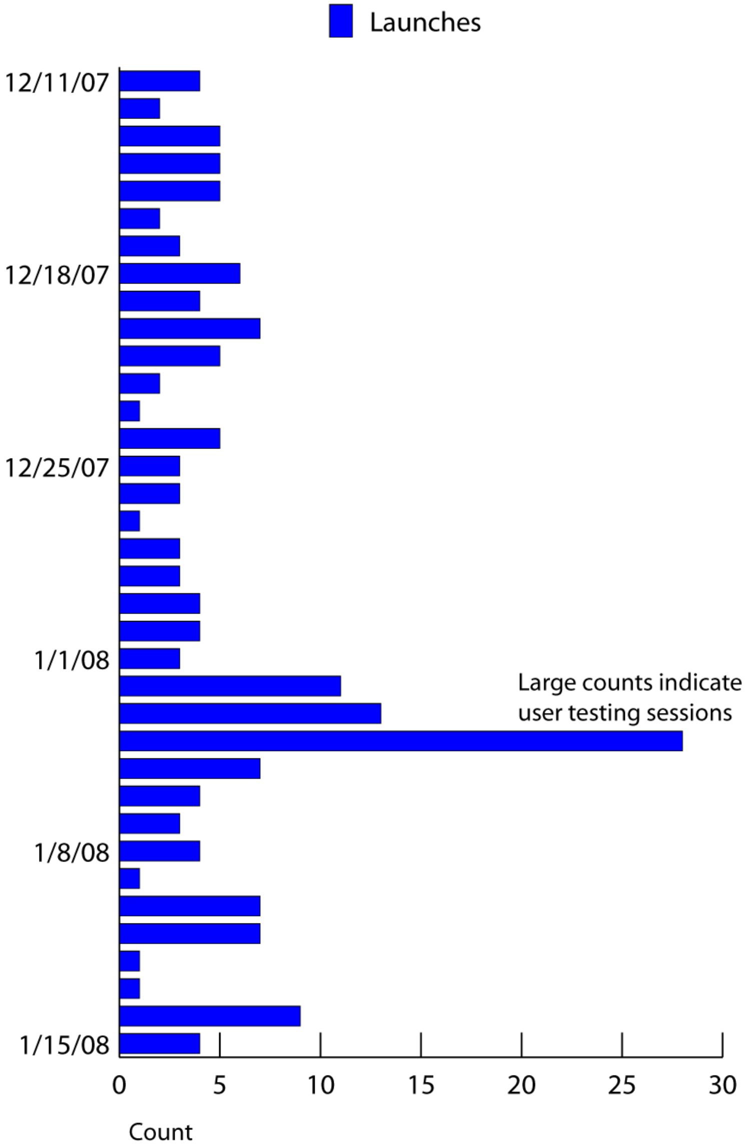
Jumbo - Dedicated Usage - 12,900 Nodes/Job



Downtimes



Jobs Running/Day



Large counts indicate user testing sessions

Moral of the story…

- Lots of faults and fault types.
- Yes it is a big problem now…
  Yes it will be a bigger problem later.
- But expert drivers get the job done!

Questions for speaker:

- What can you tell us about your fault/error situation currently?
  - What rate of errors?
  - What types of errors do you see?
  - How does one know if there is a fault?

- What are you worried about in the future?
  - What do you think is needed to help?
  - What do you think it is reasonable for apps people to do?

What are you worried about in the future?

- Silent corruption / soft errors (Michalak, Bronevetsky, …)

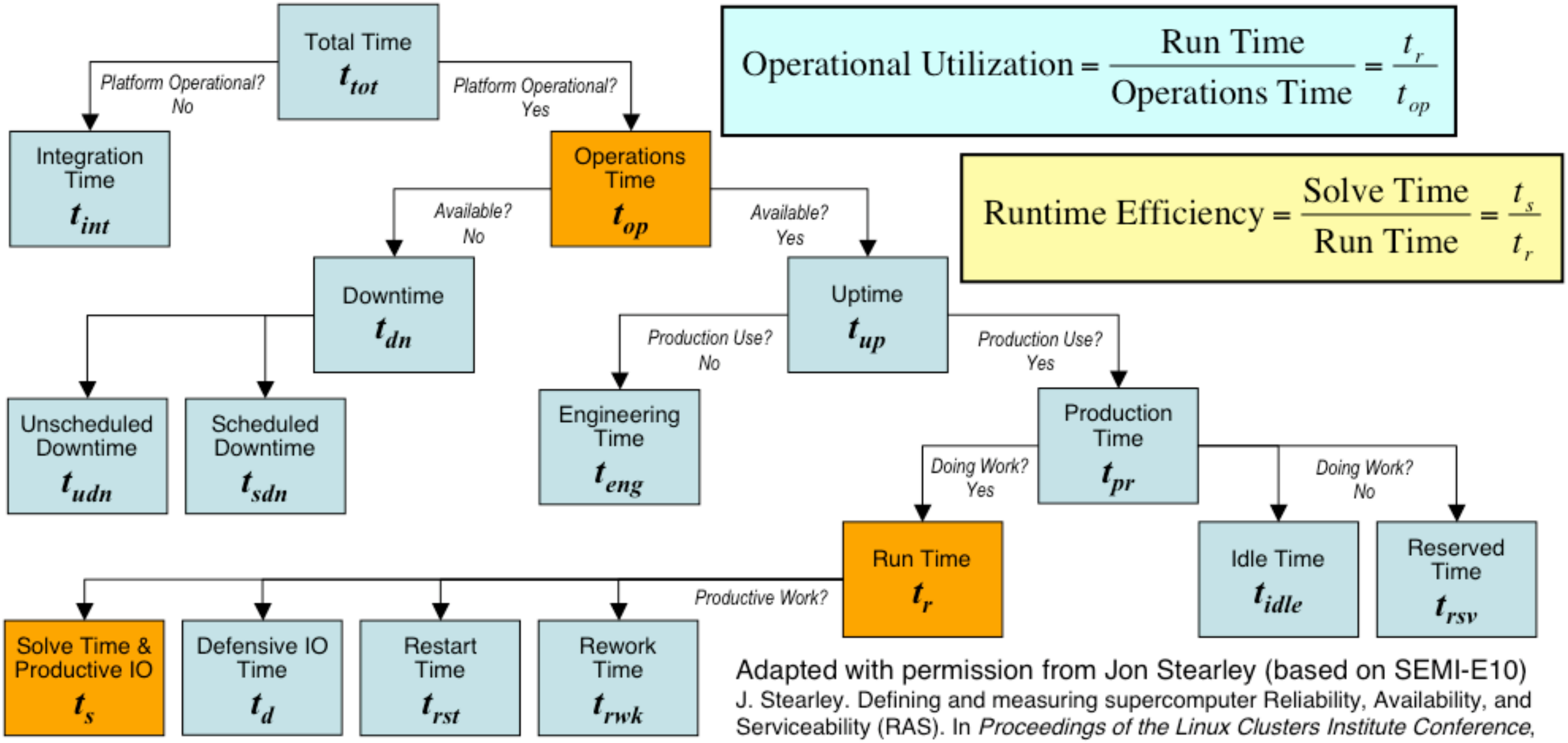
And the lack of…

What do you think is needed to help?

- **Well-defined, Standardized Metrics! (Stearley, Daly)**

A Slide From John Daly:

Defining a Productive Work Rate in Terms of How the System is Spending its Time



$$\text{Operational Utilization} = \frac{\text{Run Time}}{\text{Operations Time}} = \frac{t_r}{t_{op}}$$

$$\text{Runtime Efficiency} = \frac{\text{Solve Time}}{\text{Run Time}} = \frac{t_s}{t_r}$$

Adapted with permission from Jon Stearley (based on SEMI-E10)

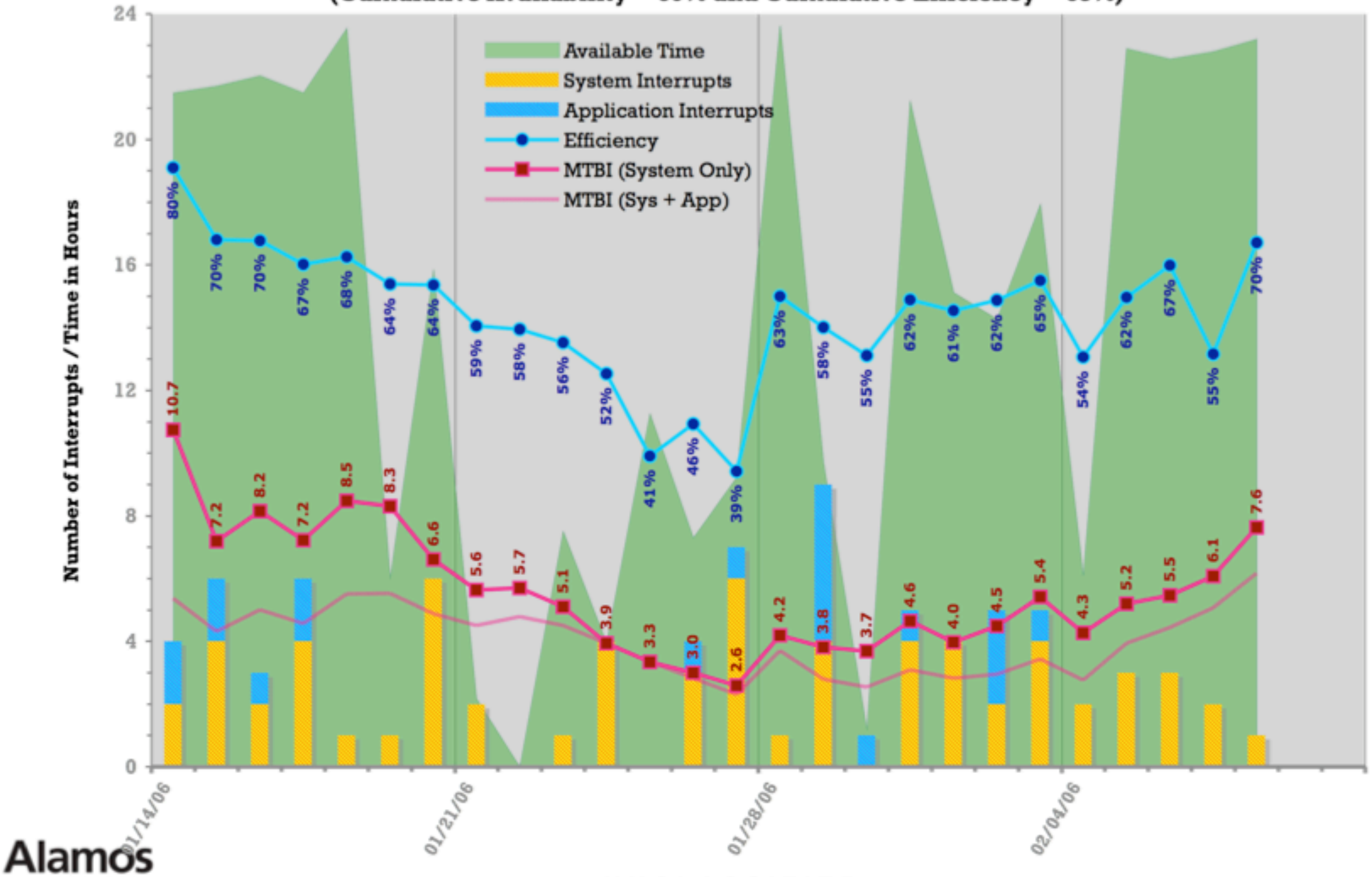
J. Stearley. Defining and measuring supercomputer Reliability, Availability, and Serviceability (RAS). In *Proceedings of the Linux Clusters Institute Conference*, 2005. See http://www.cs.sandia.gov/~jrstear/ras.

$$\text{Productive Work Rate} = \text{Efficiency} \cdot \text{Utilization}$$

A Slide From John Daly:

Operations Rate Only Tells Part of the Story: Red Storm From The Application's Perspective

Daily Availability for the 5000 Node Job with 7-Day Average MTBI and Efficiency (Cumulative Availability = 60% and Cumulative Efficiency = 63%)



What do you think is needed to help?

- Well-defined, Standardized Metrics!

- Architecture-independent integrated subsystems with useable licenses
  - Don’t try to solve everything with your monolith!
  - Eg CIFTS, OVIS, Sisyphus, TriPOD Monitoring suite, …

- Good fault reporting.
  - Don’t just exit(), log what you were trying to do!
  - Need a better reporting mechanism.
    - Systems people have systems data, apps people have apps data - we need more common ground.

Sandia
National
Laboratories

What do you think is reasonable for apps people to do?

- **Practice good fault reporting.**

- **Learn best-practices for app fault-tolerance**
  - **Metrics and checkpoint optimization (Daly)**
  - **Checkpoint compression (Gibson, Schroeder)**

- **Be aware of fault-tolerant algorithm research**
  (eg Chen et al)

- ***Collaborate* with systems people towards fault tolerance (eg Glosli et al)**
  - **You can't solve the problem alone,**
    and neither can we.

Sandia National Laboratories

Essential Feedback Loops

systems and apps, both man and machine

The End

Jon Stearley <jrstear@sandia.gov>

Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company,
for the United States Department of Energy's National Nuclear Security Administration
under contract DE-AC04-94AL85000.

- **What do all these exit codes mean?**

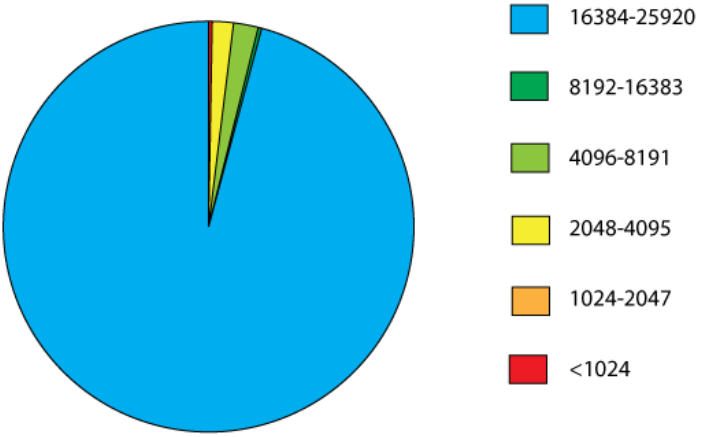
- **Which ones indicate system faults?**

- **What are correlated factors?**
  - **Username?**
  - **Application name?**
  - **Node?**
  - **Network switch? Cable?**
  - **Syslogs? App logs? …**

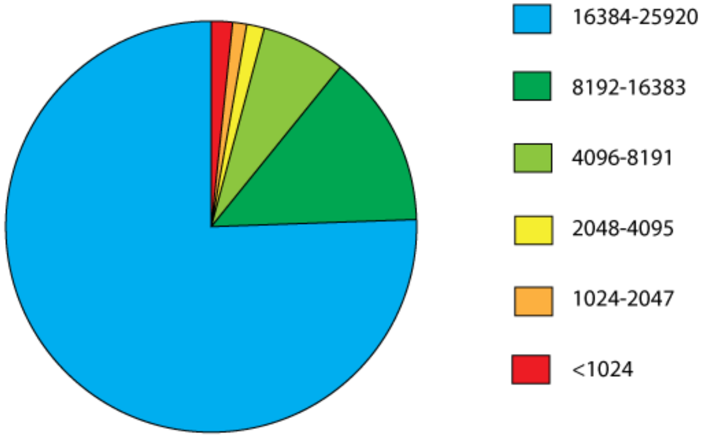
(we need better ways to investigate such questions…)

```
select count(*),exit_info from yod_accounting
group by exit_info order by count(*) desc;
count(*)     exit_info
  1396952    Yod exit code 0, signal 0
    19693    Yod exit code 115, signal 0
    14968    CLEANED UP BY RAS
     8277    Yod exit code 0, signal 2
     6962    Yod exit code 0, signal 15
     4481    Yod exit code -9, signal 0
     2735    Yod exit code 110, signal 0
     1586    Yod exit code -10, signal 0
     1428    Yod exit code 133, signal 0
     1089    Yod exit code -23, signal 0
      842    Cleaned by shutdown script
      561    Yod exit code 118, signal 0
      437    Yod exit code 27, signal 0
      433    NULL
      263    Yod exit code 10, signal 2
      241    Yod exit code -10, signal 2
      238    Yod exit code 23, signal 0
      203    Yod exit code 0, signal 1
      197    Yod exit code -23, signal 15
      185    Yod exit code -27, signal 0
      179    Yod exit code -29, signal 2
      152    Yod exit code 103, signal 0
      144    Yod exit code 132, signal 0
      139    Yod exit code 10, signal 15
      122    Yod exit code 8, signal 0
      119    Yod exit code 0, signal 13
       98    Yod exit code 120, signal 0
       96    Yod exit code -8, signal 2
       78    Yod exit code -10, signal 15
       77    Yod exit code 23, signal 15
       68    Yod exit code 9, signal 0
       60    Yod exit code -104, signal 0
       54    Yod exit code 1, signal 0
       54    Yod exit code -29, signal 15
       53    Yod exit code -23, signal 2
       49    Yod exit code 0, signal 21
       44    Yod exit code 126, signal 0
       30    Yod exit code 135, signal 0
       28    Yod exit code -29, signal 0
       25    Yod exit code 23, signal 2
       24    Yod exit code 140, signal 0
       23    Yod exit code -9, signal 2
       22    Yod exit code 111, signal 0
       20    Yod exit code -9, signal 13
       19    Yod exit code 1, signal 15
       15    Yod exit code 1, signal 2
       11    Yod exit code 8, signal 2
```

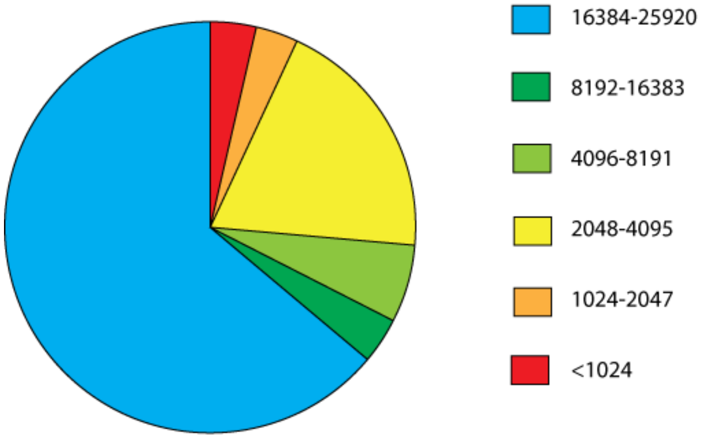
Red Storm Jumbo Mode, 12/8/07 – 3/11/08
System CPU Hours by Job Size



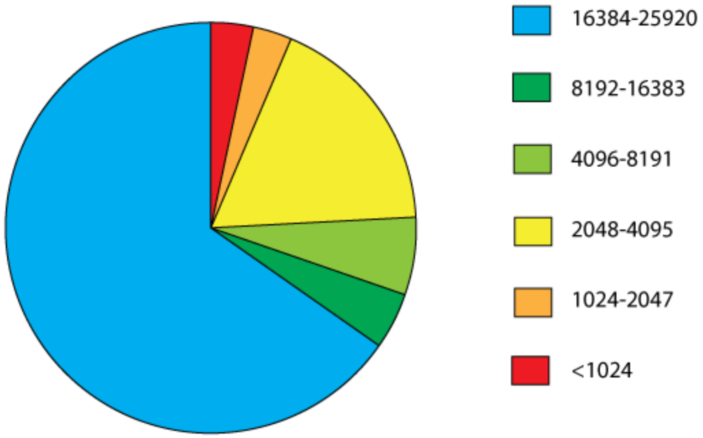
Classified
Opus Computation



Unclassified
All Computations



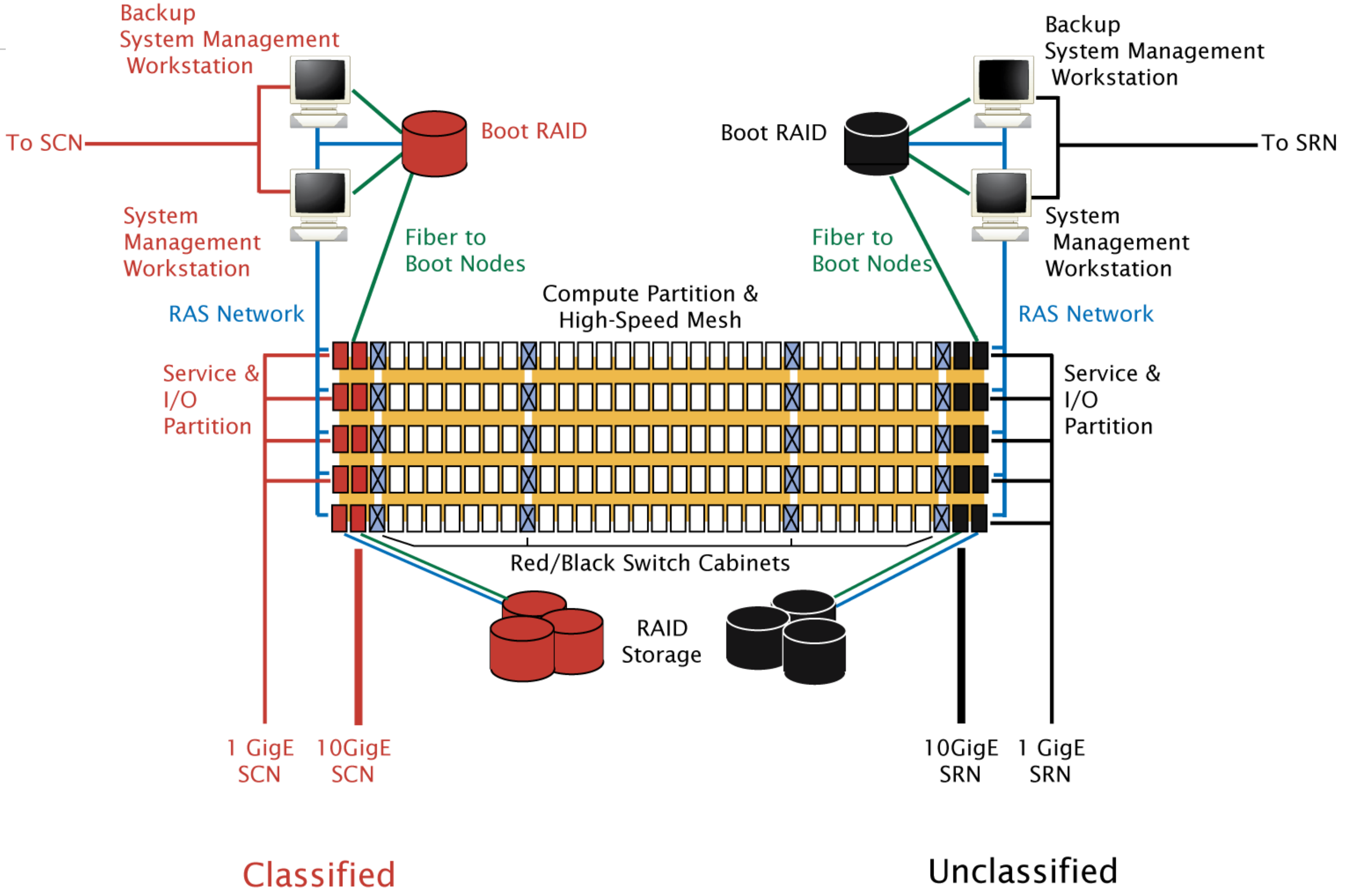
Classified
All Computations



All Computations

Sandia
National
Laboratories

Red Storm


Backup
System Management
Workstation
To SCN
System
Management
Workstation
Boot RAID
Fiber to
Boot Nodes
RAS Network
Service &
I/O
Partition
Compute Partition &
High-Speed Mesh
Boot RAID
Fiber to
Boot Nodes
Backup
System Management
Workstation
To SRN
System
Management
Workstation
RAS Network
Service &
I/O
Partition
Red/Black Switch Cabinets
RAID
Storage
1 GigE
SCN
10GigE
SCN
10GigE
SRN
1 GigE
SRN
Classified
Unclassified